DOI: 10.1111/1468-4446.12746

WILEY

SPECIAL ISSUE

Put to the test: For a new sociology of testing

Noortje Marres[1] | David Stark[1,2]

[1]University of Warwick, Coventry, United Kingdom

[2]Columbia University, New York, NY

Correspondence
Noortje Marres, Centre for Interdisciplinary Methodologies, University of Warwick, Coventry CV4 7AL, United Kingdom.
Email: n.marres@warwick.ac.uk

Funding information
H2020 European Research Council, Grant/Award Number: 695256

Abstract

In an age defined by computational innovation, testing seems to have become ubiquitous, and tests are routinely deployed as a form of governance, a marketing device, an instrument for political intervention, and an everyday practice to evaluate the self. This essay argues that something more radical is happening here than simply attempts to move tests from the laboratory into social settings. The challenge that a new sociology of testing must address is that ubiquitous testing changes the relations between science, engineering, and sociology: Engineering is today in the very stuff of where society happens. It is not that the tests of 21st-century engineering occur within a social context but that it is the very fabric of the social that is being put to the test. To understand how testing and the social relate today, we must investigate how testing operates *on* social life, through the modification of its settings. One way to clarify the difference is to say that the new forms of testing can be captured neither within the logic of the field test nor of the controlled experiment. Whereas tests once happened inside social environments, today's tests directly and deliberately modify the social environment.

> *… anything which can exist at any place and at any time occurs subject to tests imposed upon it by surroundings, which are only in part compatible and reinforcing. These surroundings test its strength and measure its endurance.*
>
> *John Dewey, 1998 (1925)*, Experience and Nature, *p. 70*[1]

1 | INTRODUCTION

"Have you been tested?" Context is everything. What counts as a test is completely different, depending on whether you are in a doctor's office, on the website of the Department of Motor Vehicles, or in one of the 12 Life in the UK Test centers for new citizens in London. "Have you taken the test?" is different from "Were you involved in the test?" for the former suggests that the test was based on a decision, whereas the answer to the latter might well be "I really don't know." We are tested. Things are tested. Models and machines are tested, and character, too. There are stress tests for banks, cardiac patients, and nuclear power plants, there are personality and citizenship tests, sound tests and screen tests, and there are tests of strength and tests of faith.

Tests—technological, psychological, pedagogical, medical, juridical, statistical, political, religious, the list goes on—are an important part and familiar form of modern society. They signal that things and people cannot really be known in general or in advance but that their attributes ultimately need to be established on a case-by-case basis. Tests have become such a familiar genre that today almost anything can be a test situation. Sometimes we recognize that testing is underway but sometimes we are unaware. The logics and grammars of testing are so readily at hand that they can be picked up and used in many fields. Furthermore, in an age of computational innovation, more and more environments in society are equipped to facilitate testing, from airport shopping malls to social media platforms.

In recent years, sociologists and scholars in cognate fields like history, anthropology, and philosophy have insisted on the importance of experiments undertaken in social environments "beyond the laboratory," describing them as sites where new forms of governance, economy and subjectivity are invented (Engels, Wentland, & Pfotenhauer, 2019; Mills & Tkaczyk, in press; Murphy, 2006; Van de Poel et al., 2017). In this essay, we argue that something more radical is happening than simply attempts to move tests from dedicated test sites into social settings like streets and social media platforms. Rather, it is the relations between test sites, on the one hand, and environments in society, on the other, that are changing. As we elaborate in the sections that follow, these recent developments pose a challenge to sociology and, in particular, to the sociology of testing.

We will propose the following: to grasp the significance of the rise to prominence of testing "beyond the laboratory,"[2] testing in society should be studied from the standpoint of their consequences, that is, *on the basis of what tests generate*. Making this argument (presented in greater detail in subsequent sections) means that our approach to testing casts a broad net. Rather than beginning with a restricted definition of tests, we proceeded from a posture of inclusion. It is the case that different fields have laid claim to, and have been defined in terms of, a specific logic of testing (on the natural and the social and political sciences, see Popper, 1961, 2002, respectively; on music and the avant-garde, see Cage, 1973; on democratic politics, see Dewey, 1954). But testing as a way of knowing, valuing, and intervening does not belong to any one social domain in particular, be it science, technology, politics, culture, or religion.

In music, a "demo" is a kind of market test. The same in TV and film where it is known as a "pilot." Pilot studies in science are preliminary tests of a model or concept. In architecture a model is a test, a kind of experiment (Yaneva, 2005). The draft of a manuscript is an internal test. Circulating a manuscript as a work in progress invites users to test for bugs, a form of academic "beta testing" (Neff & Stark, 2004). A rehearsal is a kind of test (Buchmann, Lafer, & Ruhm, 2016). Conversely, tests of emergency practices and of their management are stagings that can be thought of as a kind of rehearsal (Davis, 2007). In jazz, a jam session can operate to test along multiple dimensions (Hennion, 2003). It can be a test of new melodic, thematic, rhythmic, or choral structures. In this sense, a jam session is experimental and exploratory. It can also be a test to find a new musical partner. In that sense, it is like a try-out and can vary according to whether it tests the skilled performance of the new candidate or tests the compatibility of the existing ensemble with a new candidate (DeVeaux, 1997).

In the field of romantic attachments, a date is a test—of the invited partner but perhaps also of self. New forms of online dating provide more opportunities for population-level testing to see which presentations of self attract more (or different kinds of) responses (Ellison, Heino, & Gibbs, 2006). In 2016 Volkswagen was caught cheating on diesel emissions tests. And in 2018, the British Home Office revoked visas for tens of thousands of students who

were alleged to have cheated on an English language test, allegations based on spurious analysis of test results by the US-based private test provider (Bulman, 2019). Most tests are deliberate; but they can also be incidental, as some situations can be grasped, after the event, as, in fact, a test. Others can be a ruse: by reframing a prior action as a test, they offer a disclaimer for an earlier misstep: "It was just a test."

This recognition that testing does not belong to any particular domain or even era[3] sets us apart from other approaches in the social study of testing. In Science and Technology Studies (STS), tests and testing have long been considered to be what makes science and innovation "special": it is from the capacity to test in a laboratory that science and engineering have been said to derive their exceptional power to render invisible or distant natures observable (Latour, 1993; Mody & Lynch, 2010; Vertesi, 2015) and durably transform socio-technical arrangements in society (Callon, 1986; Laurent & Tironi, 2015; Pinch, 1993). We therefore suspend some assumptions held in these previous social studies of testing, chief among them the idea that testing "naturally" belongs to the sciences, engineering, or even to experts.

The notion that the sociology of testing should move beyond the laboratory or the field experiments of technologists was central to the work of the French sociology of critical capacities (in French: the sociology of *épreuves*, literally meaning trials). In the 1990s, this emerging school proposed that critical moments are ones in which justificatory claims, based on the principles of market, inspiration, efficiency, loyalty, fame, and the civic (Boltanski & Thévenot, 2006), are put to the test (Stark, 2011). We draw on this work as one of several points of departure. Yes, testing occurs in many sites; but it is not only about justificatory claims. And yes, as well, it is important to grasp the test as a critical moment; but it would be mistaken, as we elaborate below, to characterize tests first and foremost in terms of the forms of justification or orders of worth they elicit. Most importantly, whereas Boltanksi and Thevenot analyse social life as giving rise to trials (making comparisons to scientific and juridical trials) that yield a judgment, we argue that tests should be studied not on the basis of what they resolve but by what they generate.

Tests are generative, they stimulate further testing, not only as *experimental or test regress* in the same modality (see for example Collins, 1981, 1985) but involving *diverse modalities* (Stark, 2011; Marres, 2012) of knowing, valuing, and acting. As we shall argue in more detail below, when one studies what results from the test and not simply the "results of the test"—that is, when one examines consequences less as what is resolved than by what is further generated—then research on testing cannot remain focused on a particular testing moment but must study whether and how any given test operates in an *ecology of testing*. This is a situationalism (Knorr-Cetina, 1981) that attends to the structure of the situation (Stark, 2017).

Moreover, our intuition is that contemporary forms of testing (real-world experiments, platform-based testing, RCTs) mark a more distinctive departure than simply a movement from laboratory to field. As a starting point, think of them as testing "in society" (van de Poel, Asveld, & Mehos, 2017). And then, more radically, begin to see that, in many of the *real-world testings* examined here, tests are not just *in society* but are tests *of society*. The challenge that a new sociology of testing must address is that the very relation between science, engineering, and sociology is changing: Engineering is today in the very stuff of where society happens—and not simply because technology is embedded in or infused by the social. Instead, in our era, engineering tests the very fabric of the social. As we shall see, it is not that the tests of 21st -century engineering occur within a social context but that it is the social that is being put to the test.[4]

2 | POINTS OF DEPARTURE: THE SOCIAL STUDIES OF TESTING AND THE SOCIOLOGY OF EXPERIMENTATION

To examine how the role of testing in society is changing, and to figure out what should be the next steps in the sociology of testing, we draw on and move beyond two established sociological approaches: the social studies of testing developed by scholars in STS from the 1980s onwards, and the sociology of experimentation, which can be traced back to American pragmatist sociology of the Chicago School of the 1920s.

2.1 | The social studies of testing: Finding the social inside the test

In science and technology studies, tests have been granted special importance as sites where the inter-relatedness of science and society becomes visible (Callon, 1986; Latour, 1996; Woolgar, 1990). In the late 1980s, as scholars began to extend insights from the sociology of science to the sociology of technology, they attributed central importance to tests, viewing them as the technological equivalents of scientific experiments: Tests are to technology as experiments are to science (Pinch, 1993). Just as the Sociology of Scientific Knowledge (SSK) had done with science before, sociologists of technology studied tests to demonstrate that there is an irreducible social dimension to testing. In "From Kwajalein to Armageddon?"—a study of intercontinental ballistic missile testing between California and the Marshall Islands—Donald MacKenzie presents this now familiar, Wittgensteinian critique of epistemology:

> *… the result of most testing is accepted routinely as fact. … [N]either "logic" nor "reality" are sufficient to explain this. Other elements must be involved, and in them "the technical/scientific" is inextricably interwoven with "the social". (MacKenzie, 1989, p. 415)*

Striking about MacKenzie's article today is how he defines the "social." He views the social primarily in terms of *convention*, where the testability of any claim or proposition is "defined by the consensual practices of groups of accredited practitioners" (MacKenzie, 1989, pp. 416–417).[5] For MacKenzie, as for other SSK scholars before him, these social practices are the determinant that makes up for the indeterminacy of scientific knowledge. The task for the sociology of testing, for him, is to account for the *legitimacy* of technological testing, and the knowledge it produces, in social terms. The idea that expertise has legitimacy can today easily be doubted. In a historical moment in which the public credibility of science and technology are widely questioned, it cannot be the job of the sociology of testing to explain why they are *not* being challenged in our society.

While MacKenzie takes care to foreground the social environment in which ballistic missile testing takes place—the essay opens with a description of the islanders' situation,[6] his conception of the *social dimension of testing* makes no reference to the test environment. Nor, given his objectives, should he have done so. If the goal of the SSK project was to demonstrate the irreducible social dimension of technology testing, then it was necessary to analyze a "hard" case focusing on highly technical and abstract statistical operations performed in dedicated test sites "away from society."[7] To engage conceptually with the consequences of testing "on the ground" would get in the way of testing SSK's theory of testing. But this belief that it was necessary to study technology testing in dedicated environments at a remove from everyday environments in society came with a cost: SSK's *conception of the social*—in terms of the conventions that prop up the credibility of science and scientists—followed (at least partly) from this investment and was then reinforced by its findings. It is this particular treatment of the social in SSK studies of testing that we believe is no longer satisfactory today. Looking back at this literature from a distance, we are particularly struck by the abstract vocabulary—"convention" securing the "legitimacy" of science—relied on to characterize the social aspects of testing.

Actor-network theory (ANT) was different, in that their studies of technology testing, such as Callon's (1986) well-known study of the electric car, valued tests as sites where the power of innovation to transform society could be demonstrated *by different means* (different that is, from the political, legal, and cultural means by which sociology had typically understood social change, and social order, to be accomplished). Consequently, ANT studies of technology testing did not rely on existing sociological vocabulary, as in MacKenzie's case, to account for the social dimension of testing. Instead they advanced novel concepts, such as "enrolment," "translation," and "interest alignment," to highlight the capacity of tests to implicate social actors in the innovation process, or more precisely, to foreground the capacities of technology testing for ontological transformation. For ANT, the test enables the performative specification of a new set of connections between technical elements, social actors, objects, interests, and so on (Callon, 1984, 1986; Muniesa, 2014).[8]

While SSK and ANT had much to disagree about—ANT proposed "pay attention to the non-humans," while SSK and their successors in the social construction of technology (SCOT) program claimed "don't forget about the humans!" (Bijker & Pinch, 2012)—there was in retrospect much that they shared: both developed new kinds of explanations of the power of science and innovation by locating the social inside the test. By contrast, the sociology of testing advocated here makes the reverse move—it shows how testing and experimentalism have been extended into distinctively social environments, so that it is now society and social life itself that is subject to the technological regime of testing.[9,10]

2.2 | The sociology of experimentation: Finding testing in society

As noted, one of the pressing questions for the sociology of testing today is how to make sense of the implementation of tests and the creation of new types of test environments—smart city pilots, for example, or randomized control trials for economic development—in society (in hospitals, streets, villages, media platforms, etc.). To do so, scholars across the social sciences have in recent years turned to a sociological tradition, the Chicago School, and in particular, the classic proposition of Robert Parks to regard social environments like cities as laboratories (Gieryn, 2018; Gross, 2009; Gross & Krohn, 2005; Guggenheim, 2012). This proposition is today celebrated as a relevant precursor of contemporary efforts to create living laboratories, test beds, and experimental cities (Engels et al., 2019).[11]

In these accounts, the Chicago School stands out for its proposition that social life is already experimental in and of itself (for a discussion see also Marres, Guggenheim, & Wilkie, 2018). Building on the pragmatist philosophy of John Dewey—who had extended the notion of experiment to a wide range of activities, including personal conduct, the running of government, and democratic movements—the Chicago School proposed to study the city of Chicago, its neighborhoods and communities as laboratories. As both Thomas Gieryn and Matthias Gross note, the case for Chicago as laboratory was made on the grounds of its diversity, its social problems, and the unprecedented encounters between relative strangers (African Americans, immigrants and "hobos") in this city, who needed to figure out (try, test) how to live together (Thrasher, 1927, p. 488; Tolman, 1902, p. 116). However, others offered general justifications for the approach. As Chicago sociologist Vivien Palmer (1928) argued, "human beings are everywhere so continually performing their own experiments in group life that the investigator can always find social experiments of many kinds in progress: a systematic, contemporary, observation of these yields significant facts" (p. 8).

The intellectual project of the sociology of experimentation is, thus, in many ways the opposite of 1980s' STS studies of testing: whereas the latter finds the social inside testing, the former finds testing inside the social. The key insights of the Chicago School that social practices increasingly present themselves as experiments (Gross & Krohn, 2005, p. 80) and that sociologists should take advantage of this experimentation by everyday social actors in society (Gieryn, 2018, p. 15) was initially ignored by SSK and ANT studies of testing, where the focus was on hard science and innovation in environments at a distance from the familiar settings of everyday life. Conversely, for a contemporary sociologist like Matthias Gross studying the Chicago School, the experimentality of social life thematized by Park and colleagues has very little to do with traditional notions of science. For him, social experimentation is "completely different" from social science experiments; to treat society as an experiment is to adopt a "sociological perspective [that] has got nothing to do with the idea of sociologists as experimenters in white coats" (Gross & Krohn, 2005, p. 80). Most pointedly in their mirrored opposition, whereas SSK scholar Harry Collins regarded the scientific experiment as a template for all forms of cultural production (Collins, 1985, p. 18, 165),[12] Chicago School advocate Martin Gross argues that "experiments in the real world are, in a sense, the real and true experiments, and the laboratory ideal is a special variation of it" (Gross, 2009, p. 90).[13]

Our task here is not to adjudicate among claims about which are the "real and true experiments" upon which other kinds of tests should be modeled. Instead, it is to understand the achievements and limitations of these two approaches for analyzing testing today. If SSK persuasively demonstrated that technological testing is suffused

with social processes, its abstract conception of the social restricted its choice of research settings to specific sites away from mundane society—thus, bringing in the "social" but, in effect, avoiding the societal. If the rediscovery of the Chicago School by scholars of testing decidedly moves our attention onto this otherwise neglected terrain, its notion of the self-sufficiency of social experimentation—that social life is already experimental in and of itself—ignores how expert-led experiments deliberately introduce something new into social life. The proliferation of tests across society is today intimately connected to the organizational, technical, material *modifications of society*, often by computational means—modifications that render social life observable, analyzable, and influenceable *by a variety of actors* (Marres, 2017). In this context, the primary question is not whether social practices can be *defined* as experiments, but how social environments—and possibly social life itself—are *modified* so as to enable experimental operations by a range of different kinds of actors, and putting these actors into a determinate, often unequal, relation.

To study these contemporary tests in and of society, we believe it is necessary to let go of a number of assumptions that, despite their differences, the two approaches share, chief among them the underlying assumption that expert-led testing and social experimentation are best understood when studied in isolation from one another. In our view, it is the very division between the social studies of testing and a sociology of experimentation that risks rendering invisible the testing situations that the sociology of testing should elucidate. It is not just that in these situations, *both* engineers and scientists, on the one hand, and social actors, on the other hand, tend to be implicated. As or more important is that *a diversity of modes of testing* and experimentation are detectable in them: scientific, political, aesthetic, commercial. In recent street trials of autonomous vehicles, engineering tests do double duty as public engagement experiments, and these trials are equally enlisted to conduct activist and creative interventions into street environments, as when the artist James Bridle trapped an autonomous car by laying a salt circle on the surrounding road.[14] It is not just that engineering tests and social experimentation happen in the same social space, the street; rather, environments in society are being modified by a variety of test subjects and agents, in such away that diverse forms of experimentation pose a challenge to one another.

This certainly does *not* mean that the differences between technological testing and social experimentation are today being dissolved, but it does follow that if we study these diverse forms of testing in isolation from one another, we are unlikely to understand how society is transformed by means of testing. We therefore argue for a *new sociology of testing*. But to make that case, we need to do more than push back against existing theoretical frameworks. We need materials to think with, and so we offer two brief accounts, examples of testing in two historical periods, the first from the mid-1950s, the second from our current time some 60 years later. In methodological terms these are not *cases*, and although they involve tests, they are emphatically not offered as *tests* of competing theories. We make our argument in steps. The first example provides an opportunity to point to changes in research strategy that would need to be made if the sociology of testing is to account for testing *in society*. The second example points to changes in infrastructures and practices in society that the sociology of testing will need to take into account in a new era of testing.

3 | FOR A NEW SOCIOLOGY OF TESTING I: STUDYING TESTING IN SOCIETY

3.1 | 1956: The fallout of nuclear testing

In March 1959, *Consumer Reports*, a monthly publication of the Consumer Union, a nonprofit consumer advocacy organization, published a report of their product testing that found radioactive isotopes in milk. Readers of the US magazine would typically consult its pages for reports on the testing of products such as cars, tires, or household appliances. But in this report on "fresh clean milk, which looks and tastes just like it always did" they learned, many for the first time, about strontium-90. A "toxic substance known to accumulate in human bone," this "unseen

contaminant" was a health hazard caused by radioactive fallout (Consumer Reports, 1959, p. 103). In testing for the adverse effects of radiological food contamination, the product testing of milk traced a line back to an earlier (and, at the time, ongoing) set of tests—atmospheric nuclear testing.

The first test of a nuclear explosive device, named "Trinity" by J. Robert Oppenheimer, took place in the New Mexico desert at 5:30 a.m. on July 16, 1945. Surface level radiological monitoring was limited to 200 miles from ground zero. But the effects of the Trinity test were incidentally measured far beyond the distances accounted for in the army's surveys. In August of that same year, batches of photographic film belonging to the Eastman Kodak Company in Rochester, New York, began to show inexplicable areas of fogging (Stannard, 1988, p. 885). A physicist at Kodak's Research Laboratory traced the phenomenon to the interleaving paper used to separate photographic films in their packaging. It turned out that this interleaving paper was produced from corn husks, which were traced to a farm in Indiana, some 1,300 miles from the Trinity test site (Bruno, 2003, p. 140). The Kodak physicist concluded that the corn husks were contaminated by radioactive fallout, and eventually published these findings in 1949.

With the onset of the Cold War, atmospheric nuclear testing accelerated in both the United States and the Soviet Union, reaching its peak in 1961–1962 when about 250 megatons were detonated in the atmosphere (UNSCEAR, 2000, p. 160). In 1949 the US Atomic Energy Commission (AEC) launched a parallel test—an investigation into the "long term, widespread hazards" of nuclear testing. The project, code-named Gabriel, was followed by another, Sunshine, in 1953. Analyzing how radioactive isotopes such as strontium-90 found their way into "air, water, soil, plants, animals, and humans" (Bruno, 2003, p. 240), Gabriel and Sunshine were, at one level, tests of testing. But because these investigations hinged on the question of "how many bombs could be detonated before reaching doomsday level?" they were framed as an "essential preparation for a nuclear war" (Bruno, 2003, p. 243). In an important sense, Gabriel and Sunshine were *stress tests*. That is, they were not only conducted to test the effects of nuclear testing, but were also conducted as a test about a catastrophic event—all out nuclear war—that had not actually occurred but could be imagined or projected into the future.

For these projects, the AEC collected data on the pathways of contaminants not only in the food chain but also in the upper atmosphere and in the oceans. Incidentally, the Gabriel and Sunshine tests greatly improved the general understanding of circulation on the planet. It was through such studies, for example, that "atmospheric currents such as the jet stream were detected" (Bruno, 2003, p. 245). Every cloud, even a mushroom cloud, it seems, can have a silver lining.

But the Air Force and the AEC were not the only agencies conducting tests. Following the declassification and publication in 1957 of the Project Sunshine findings about the presence of strontium-90 in nuclear fallout, the US Public Health Service began to monitor levels of radioactivity in air, water, and milk samples in 1958 (Watkins, 2001, p. 301). The real challenge for health officials was to find out how much of the radioactive isotope was getting into bodies. Established procedures called for bone samples which were difficult to get in large numbers. The answer was not discovered by AEC commission but by the "Baby Tooth Survey"—one of the most politically consequential test of nuclear testing conducted at the time—announced in December 1958 by the Greater St Louis Committee for Nuclear Information (CNI), a partnership of citizen and scientific activists, including the environmentalist pioneer Barry Commoner.

CNI called on children and parents to send in baby teeth, along with a 3×5 card with basic information about the donors such as birthdate, where they resided, and whether they were breast-fed or bottle-fed. Mobilized by dentists, churches, libraries, and schools, local residents responded eagerly. St Louis was an ideal site for the survey: the Public Health Service tests in nine US cities had found that the Midwestern city topped the list for having the highest levels of strontium-90 in milk, tainted by dairy cows grazing on grass contaminated by nuclear fallout. Under the direction of a local physician, Louise Reiss, and cataloged by volunteers of the Women's Auxiliary of the St. Louis Dental Society, the project collected almost 15,000 baby teeth in its first year (and more than 300,000 by the time the project ended in 1970). Reiss reported the findings of the survey in an article she published in *Science* in 1961: Children born in 1954 had four times as much strontium-90 in their teeth as those born in 1951

(Reiss, 1961, p. 1670). That study received widespread media attention and is credited as an important moment in the efforts to bring about the 1963 Limited Nuclear Test Ban treaty as acknowledged by President John Kennedy in a personal phone call to Louise Reiss (Watkins, 2001, p. 304).

3.2 | Lessons for the sociology of testing, step 1

We draw three major lessons from this account of the societal fallout from nuclear testing.

3.2.1 | Tests are generative and may give rise to an ecology of testing

Tests, we argue, should be evaluated not only on the basis of their validity or by what they resolve but equally by what they generate. For the sociology of testing, more important than the "test results" is *what results from the test*.

Tests are generative, in the first instance, because they provoke further testing. As we saw in the nuclear testing example, what began as the test of a weapon spawned tests of the atmosphere, of ocean currents, of planetary resilience, strontium-90 in baby teeth and the credibility of US federal agencies. More important, as we shall also see elaborated in the research findings in the articles of this special issue, the further tests that result from testing need not be of the same modality nor in the same register. Not just a more refined test of the test, and not a simple contagion like the spread of a given bacteria, the tests that proliferate are more like different species—technological, personal, political, artistic, and so on.

Our proposed perspective directs attention to a different phenomenon from the one that preoccupied 20th-century philosophers and sociologists of science and technology from Pierre Duhem to Harold Garfinkel as well as Donald Mackenzie noted above: The experimenter's regress. According to the latter relativist proposition, *the results of any test are inconclusive*, and therefore, in principle, can always be contested in further tests. We point to something different: A given test tends to unfold amidst other tests, and often these are *different kinds* of tests. It follows that one of the principles of the sociology of testing would be that tests should be examined within an *ecology of testing*.

When viewed in this way, we can understand why tests are not only inconclusive but also why they can be unsettling. A given test often does not resolve, and the more that it takes place within an ecology of tests operating in different modalities, the less the likelihood of clear resolution. It is the possibility for shifts and dissonant connections between different registers of testing that gives rise to *a distinctive type of ambiguity*, or more precisely, undecidability. Within such a context, testing raises the determinate possibility of change in several different directions. In such a situation contestation is not simply proposing an alternative way to measure some attribute but pointing to a different value altogether along a very different dimension and according to another accountability. This is why tests have the capacity to unsettle. And it is for the same reason that they can be productive without producing resolution (Stark, 2011).

When we look this way at the consequences of testing, we see that tests that involve technical devices can give rise to tests of social relations. This is a principal finding of Joan Robinson's (this volume) research on the home pregnancy test, "What the pregnancy test is testing." Robinson's case is a textbook example of testing "in the wild," as the pregnancy test moved out of the laboratory and "came home from the hospital."[15] The question now addressed: What is tested in the home pregnancy test? The pregnancy test is a technology of discovery. With it a woman can discover, without the mediation of medical professionals, whether she is pregnant. Robinson's sociological insight is to situate the home pregnancy test in the context of social relationships. While revealing a new relation between a woman and her body, the home pregnancy test has the potential to requalify relations between the woman and her partner, her friends, her in-laws, her high school swimming coach, her employer, and others.

Robinson demonstrates that, within such an ecology of testing, the pregnancy test (and not the simple yes/no fact about pregnancy) can put social relations to the test.

3.2.2 | Testing in society takes place at multiple sites, on differing scales, provoking emergent entities and sometimes deploying intimate relations

In our example of testing at mid-20th century we saw that the ecology of testing finds tests taking place in many sites and settings. On the American side, the nuclear weapons tests themselves took place primarily in Nevada and in and over the Pacific, including on the Bikini Atoll, a coral reef in Marshall Islands. But the literal fallout of the tests spanned the globe; and the figurative fallout led to tests in an upstate New York commercial laboratory and in a cornfield in Indiana, tests of soil samples from the dairy farms near 10 major American cities and tests of models of nuclear apocalypse. The "Pacific Proving Grounds" (as the nuclear tests site were named by the US Government) might have been far distant from Washington, DC, but they were certainly not isolated. Nor are the different kinds of tests isolated from each other. Yet, as connected, they were not associated as parts of a coordinated, encompassing test.

In addition to being multi-sited, the new sociology of testing must also be alert to processes whereby testing can be conducted by human entities not at all anticipated in the initial tests (employees at Eastman Kodak, from our example) involving non-human mediators at steps removed (e.g., corn husks and baby teeth as silent witness). In dynamic test environments, moreover, testing enables the activation of unexpected entities (in our example, an association bringing together teachers, dentists, and environmental biologists).

A given nuclear explosion can be measured on various scales, perhaps there is one that goes from a stick of dynamite to a solar storm. Any given test is likely to have many data points, but the total number of atmospheric tests as counted by the Arms Control Association (2019) has been 528. As mentioned, the Greater St. Louis Committee for Nuclear Information eventually collected over 300,000 baby teeth. Whether small or large in the number of data points, tests can involve the extraction of intimate stuff from family life. (children giving up the tooth fairy got a button exclaiming "I GAVE MY TOOTH TO SCIENCE"). The cold calculations of the stress tests to estimate the survival of the human species in the aftermath of nuclear war were conducted on Scale 1—the planet.

3.2.3 | The relation of testing agents and tested subjects is unstable and to an extent reversible, sometimes leading to a test ban

American citizens, Soviet citizens, human beings living anywhere on the earth from Nova Scotia to New Zealand, from Beijing to Buenos Aires, all were involuntarily subjected to the fallout of nuclear testing. Although the initial object of the atmospheric nuclear tests was the weapons system, with further testing of the effects of nuclear testing, humans became research subjects. Some citizens resisted nuclear testing and nuclear weapons absolutely, most famously the Campaign for Nuclear Disarmament in the UK, whose semaphore symbol schematically became the international peace symbol. But, as our example shows, in addition to publicly demonstrating, citizens carried out their own tests, the results of which were offered as demonstrations that strontium-90 was being absorbed into infant bodies. That is, instead of the ANT dichotomy of compliant versus unruly, these subjects were *contesting*. The St Louis moms put the Atomic Energy Commission to the test.

No one volunteered to be the subject of nuclear testing. Most significantly, no one could opt out. No single person could choose to be individually exempted from being involved in atmospheric tests. For any person to be protected from the results of nuclear testing, everyone on the planet needed to be protected. The answer: A test ban.

4 | FOR A NEW SOCIOLOGY OF TESTING II: STUDYING CONTINUOUS, UBIQUITOUS TESTING

4.1 | 2016: Fallout from online personality testing

In 2013, scientists at Cambridge University—notably Michal Kosinski and David Stillwell at the Psychometrics Centre—developed a third-party Facebook app, mypersonality.org. For this online personality test, users answer some questions and receive their scores on a five-factor model of personality, the so-called "Big Five," or "OCEAN" model (for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; see John & Srivastava, 1998). Various prior studies had shown that personality traits can predict patterns of activity such as consumer purchases, sexual orientation, or political behavior. Of interest to the Cambridge University researchers was the following question: Could these traits be found to correlate with online behavior such as "liking?" If so, one could use data gathered through social media to make inferences about personality (even without the user taking a personality test).

On the basis of their research, Kosinski and colleagues answered this question in the affirmative, publishing a number of scientific articles positing that the analysis of social media data captured by means of online tests could predict private attributes such as sexual orientation and political preference (Bacharach, Grepel, Kohli, Kosinski, & Stillwell, 2014; Kosinski, Stillwell, & Graepel, 2013; Lambiotte & Kosinski, 2012). While the study attracted significant media attention because of its findings, including headlines such as "Gay? Conservative? High IQ? Your Facebook 'likes' can reveal traits" (Boyle, 2013), Kosinski and colleagues themselves addressed publics as testing subjects, as they publicized this research through the website, youarewhatyoulike.com, where "you can sample the method for yourself," as NBC put it (Boyle, 2013). Reporting of related Facebook-based experiments highlighted the experimental potential of their insights: taking advantage of the interactive features of online platforms, it became possible to seamlessly *move from the analysis* of user attributes to *targeted behavioral intervention*, promising to render "real-world behaviour … amenable to online interventions," as computational social science colleagues put it in *Nature* (Bond et al., 2012).

The scaling up of this experiment equally depended on the roll-out of personality testing online. In Spring 2014, Aleksander Kogan created a Facebook app "thisisyourdigitallife," using Amazon's Mechanical Turk to pay users one dollar for downloading its app and completing its incorporated personality questionnaire. Approximately 270,000 individuals downloaded thisisyourdigitallife, and through Facebook's third-party APL's this allowed Kogan to harvest tens of millions (50–87 million) of Facebook profiles. Subsequently it was revealed that the company Cambridge Analytica used Kogan's app to collect and analyze data for political campaign purposes (Cadwalladr & Graham-Harrison, 2018).

While *The Guardian* journalists who broke this story became the poster-persons of the public exposure of what has become known as the Cambridge Analytica scandal, their reporting critically relied on independent counter-expertise, *and counter-testing* of Facebook's platform settings, in the effort to experimentally determine its treatment of personal data. In December 2016, Pierre-Olivier Dehaye, a Belgian mathematician with a long-standing interest in the use of personal data by tech companies, began researching the role of Facebook-based personality testing in political campaigning (Dehaye, 2017). He also reported on the concerns that Kosinski himself has about the efforts of Kogan and associates "which was trying to purchase data obtained via his team's Facebook-linked MyPersonality psychology app for reasons it would not reveal." From early 2017 onwards, Dehaye began methodically sending data access requests to Facebook to investigate Facebook personal data processes and practices, in particular, how it handles information on custom audiences, and Pixel data—that is, web browsing tracking-data (Ausloos, 2018).

In February 2017, Michal Kosinski—by then an associate professor of organizational behaviour at Stanford Business School—published new research demonstrating that it is possible to predict sexual orientation from facial images scraped from data websites using neural networks (Wan & Kosinski, 2017). Dubbed an AI gaydar, this

research also attracted significant media attention. This claim was put to the test in an especially well-publicized counter-study of the "AI gaydar," conducted by the Princeton psychologist Alex Todorov, together with research scientists at Google, Margaret Mitchell, and Blaise Agüera y Arcas. They used Mechanical Turk to organize a survey of 8,000 Americans that asked 77 yes/no questions such as "Do you wear eyeshadow?", "Do you wear glasses?", and "Do you have a beard?" as well as "Are you same sex attracted?" Comparing the results of this survey with an analysis of portrait photography on dating websites, they demonstrated that Kosinski's test reveals not sexual orientation but the photographic conventions observed in dating websites, that is, social stereotypes (Agüera y Arcas, Todorov, & Mitchell, 2018).

4.2 | Lessons for the sociology of testing: Step 2

What can we learn from our second account of testing? There is, of course, at least one major way in which the fallout from psychological testing (as an example of new forms of testing in the online setting) is similar to nuclear testing: in each case, millions learned that they were the unwitting subjects in a series of tests. We focus here on three ways in which they might be different.

4.2.1 | Engineers have moved into the social

The mid-century nuclear tests affected intimate relations in society—the radioactive particles released by atmospheric testing entered the bones and teeth of children across the country. But it required the intervention of citizen activists who mobilized intimate relations (by way of baby teeth and mothers' milk) to demonstrate these effects. Today, tests by scientists and engineers are purposefully designed to extract and deploy intimate information to deliberately act on populations and intervene in society.

For the engineers of nuclear testing then, human beings—when considered at all—were simply part of the environment in which their tests were being conducted. For the new engineers of big data testing today, the behavior of human beings is the object to be tested. Even more, for research teams at Google, Microsoft, Airbnb, Uber, Facebook, and other firms, it is less that they are conducting psychological tests *in a social environment* but that the social environment is itself the object of testing. "Uber Engineering" celebrates computational social science, and its counterparts at Facebook, Twitter, and elsewhere do the same. Some of these teams are headed by researchers with PhDs in the social sciences, including those with titles like "Lead Trust Scientist," but these research teams are more likely to be led by a "Data Scientist" with degrees not only in informatics, computer science, or network science but also other disciplines ranging from anthropology to physics. What is a social relation? What is trust? What network structures of communication erode or promote trust? In many of the research teams of platform capitalism, you would not be laughed out of the room for posing such questions and offering ways to operationalize them as formulations to be tested—precisely because these concepts deeply inform the data infrastructures that have been rolled out across environments in society and are instrumental to the value propositions of these companies.

4.2.2 | The sites and the logics of testing have changed

Social engineering, of course, is not new. Social engineers (the term dates from the end of the 19th century) have long attempted to persuade or influence voter and consumer behavior (for an insightful overview see Wu, 2017), and for this they conducted research involving tests. What is new, we suggest, is that the sites and the logics of testing have changed. In the past, the test situation was a moment that was spatially, temporally, and

infrastructurally marked off from the ongoing flow of life. Focus groups, for example, brought research subjects into a laboratory setting (Lezaun, Muniesa, & Vikkelsø, 2013). The use of field tests in marketing research (for example, to determine the effects of product placement on lower shelves versus upper shelves in the supermarket aisle, see Cochoy, 2004) entails recognition that the social environment matters, but test data were analyzed away from the field site, and insofar as the analysis informed intervention, this involved a complex chain of translations, passing through various departments, before any intervention could be made on the shopfloor (changes in product placement, for example). Online environments change this because the underlying data architectures provide for data capture, analysis, and feedback across computational networks, enabling continuous modification of user interfaces, at the front end, based on insights gathered at the back end, with it often being opaque to users which of these operations happen in which settings, inside one's device, in data centers, or somewhere else yet.

But the difference, we suggest, is even more radical because the test situation also need not be *temporally* marked off from the flow of everyday life. Online testing, as in the case of A/B testing of search engine interfaces with different segments of users (Seaver, 2019) is ubiquitous, continuous, and seamless. In principle, in the digital setting—and, increasingly, in digitally equipped settings like a mall or a smart street—anything can be a test situation, any data point can be data for a test of some kind. Because digitally traced behavior (including something so quotidian as a keystroke) can be recorded in a database, and because the very environment where behavior is monitored is equipped to influence behavior, there is the ubiquitous possibility for any subject at any point in time of being experimentally represented and intervened upon. In the digital society today, the measurement of social life is performed as part of the conduct of social life, as every click and retweet generates *at the same time* a social action and a data point (Marres, 2017).

Moreover, social life is tested as part of social life: it is not just represented, but methodically intervened upon, by various actors involved, except that some actors—such as those with the capacity to modify interface settings—occupy an infrastructurally speaking more privileged position to undertake such tests than others. Which is to say, even as testing today extends across environments in society and centers of calculation, it remains very much the case that only some actors can move between these different settings. Testing in society puts social actors in a determinate, unequal relation. As we elaborate in the concluding section, in the real-world test environment as curated by engineers, users can tweak the "settings" on their user interface, but they cannot control, nor are they knowledgeable about, the broader systemic settings of the social media platforms and digitally enhanced environments in which they are participating.

The blurring of lines between the test situation and the world to which the test is meant to refer, is what in our view is being invoked by terms like "smart," "real-world," and "living lab." It is not confined to online environments, but can be found in a variety of technology-intensive settings, for example in the medical field. The medical world was once divided into two domains, the biomedical realm of research and the clinical realm of care—in other words, between *testing* and *treatment*. Because of the uptake of new, data-intensive techniques in experimental protocols and routine treatment protocols, for example, in oncology, "care increasingly displays experimental features" (Cambrosio, Keating, Vignola-Gagné, Besle, & Bourret, 2018, p. 210). With the development of "point-of-care" testing for personalized treatment, testing is folded *into* treatment itself (Beitelshees, 2012). This form of testing, like genetic signature tests, for example, creates

> *contexts where the frontiers between clinical and research activities are constantly moving. Indeed, both activities are often carried out in the same arenas, or are interconnected, to the extent that the actors concerned find their boundaries difficult to draw. (Rabehariosa & Bouret, 2009, p. 697)*

In some clinical trials research protocols can remain open "on an ongoing basis, so to speak *sine die*" even after publication of trial results, leading some researchers to gesture to "an infinite protocol that probably foreshadows a new way of performing clinical trials" (Cambrosio et al., 2018, p. 216).

4.2.3 | Counter-testing is difficult and dominated by experts

The revelations about personal data extraction and opinion manipulation by social media platforms and platform-based social science were in part the result of counter-studies by individual experts like Olivier Dehaye. However, these revelations did not only result in public criticism of the tech sector, they equally had the paradoxical effect of making independent or critical research on social media more difficult. After the scandal broke, Facebook changed the settings of its Application Programming Interface and app review procedures, with the result that third-party applications including several social research apps, like Bernard Rieder's Netvizz, stopped working.[16] Around the same time, a group of academic researchers supported by the not-for-profit Mozilla Foundation, criticized Facebook for not doing enough to support rigorous social science, saying its tools "do … not satisfy basic criteria for validity and reliability."[13] Even if independent Facebook research—including counter tests—helped expose the harmful fallout of social media testing, it equally had a backlash effect: Facebook's subsequent interventions made counter-testing more difficult.

It is too early to say, but this challenging environment for counter-testing may yet change: The General Data Protection Regulation (GDPR) aims to legislate out of existence the original social media business model, which is based on extraction of personal information without the data subject as transaction partner (Bellanova & González Fuster, 2018). This equally threatens the test logic exemplified by personality testing, which depends on an undemanding informed consent regime to be in place. Are we on the road to a new test ban?

Recall that no individual could opt out of nuclear testing and that this elementary fact was one of the factors leading to a test ban. By contrast, the possibility to opt out—or in—is a notorious grey area in digital media environments (Gerlitz & Helmond, 2013): being listed in a user's contact book—or passing through a smart street—may be all that is needed for someone to be enlisted as a test subject. Efforts are being made to restore the possibility of opting out of online experiments, including through the GDPR, which is designed partly for this purpose. However, it is not clear if the rights-based framework of the GDPR is even enforceable: it has not yet been put to the test, as implementation has only just begun. Furthermore, it is not clear that today's online personality tests are met with the level and intensity of public disapproval that would be required to mobilize sufficient support for a test ban. The intended effect of the GDPR appears to be to produce a market for personal data services, and potentially, testing. Is personality testing ban-able? At the moment it does not look likely. Is it more likely that, rather than an outright ban, we will see a zoning of test environments in society, with low and high thresholds of testability for different types of settings, say a premium app versus Facebook; an upmarket eco-resort versus Disneyworld?

5 | FROM TESTS IN SETTINGS TO TESTING SETTINGS

We have argued that it is not enough for sociology to study testing "in society." True, testing has moved from the laboratory into society, insofar as (a) testing methodologies are today prominent in the social environment, whether it is in the form of street trials of intelligent vehicles conducted (Marres, this volume) or citizen tests conducted as part of immigration procedures by the UK home office and the Dutch border agency (Schinkel, this volume), and (b) testing methodologies developed in science and engineering are today applied to social phenomena, as in the development RCTs studied by Luciana de Souza (this volume), and the social credit system in China (Bach, this volume). But, in these and other cases, something even yet more radical than a move "beyond the laboratory" is taking place. We state it now concisely and then elaborate for clarity: Whereas we traditionally think about testing taking place *within a setting*, today's engineers are *testing the settings*. We used to think that, to understand the relation between tests and society, we must attend to the wider contexts in which tests are conducted. But to understand how testing and the social relate today, we must investigate how testing operates *on* social life, through the modification of its settings. One way to clarify the difference is to say that the new forms of testing can be

captured neither within the logic of the field test nor of the controlled experiment. Whereas tests once happened inside social environments, today's tests directly and deliberately modify the environment.

Such modifications of the settings in which social life unfold happen most clearly in computational and computationally inflected environments: on social media platforms, or inside transport environments, for example, with systems in place to monitor, analyze, and regulate movement. But settings in society may also be turned into test environments by the insertion of a mere device into existing social environments, say, a Facebook button on a coffee shop's wifi page, a "smart" traffic light onto a street, or an "intelligent vehicle" with a pedestrian detection feature passing in the street that passes its observation to an automotive data center. In this regard, the critical distinction is that between the idea that testing occurs inside the social environment (a claim that still conforms to the logic of the "field test") and the idea that testing today involves the very modification of social environments.

Take the navigation app Waze. This app relies on mobile data that its users continuously send back to determine the properties of current traffic flow, and to this end this app deploys some of its users as probes: when a traffic jam is forming, most users will get routed around the trouble spot, but some will be deliberately sent into the congested area in order to measure the properties of this disruption, effectively serving as test subjects. These "test settings" of Waze do not just imply that individual settings (give me the fastest route) may be overridden by the app, limiting the agency of users, or at least delaying their arrival time. More important for our purpose, is that by using Waze users willfully subject themselves to experimental settings: not only may a user's settings be updated to "test subject" status without the user knowing, any Waze user is in principle subject to test conditions that may be changed at any point depending on the system needs.

These types of experimental devices—which we can think of as generating *total test environments*[17]—differ both from field tests and from classical, controlled laboratory experiments. We address these in turn. In the case of the field test, the aim is to find an appropriate setting which has the required properties so that a given phenomenon can be observed there (Morgan, 2013). The concurrent expectation is that the increase of complexity and authenticity of what is observed in a field test will inevitably be traded off against a decrease in control over the phenomenon to be studied. Real-world experiments do not escape this tradeoff so much as render it irrelevant—because its logic and its goals are different, at least where testing on social phenomena is concerned. Here, the objective is not to observe an existing social phenomenon—to secure its authenticity. In place of *finding* (selecting, choosing) a setting or sampling among several settings, the operations that produce today's total test environment consist of minor modifications in the environments in society so as to render *the setting capable of data capture, analysis, and feed-back*—that is, to equip it as a test environment, to enable representation *and* intervention—even if aspirationally—on a more or less durable basis.[18]

It is not technically correct to say, in this case, that the test happens inside a given environment in society, insofar as the whole point—or at least a major point—of the test is to modify that environment, to establish connections *between* the field and the laboratory, between entities making up the street environment and the data center, so that it becomes possible to dynamically adjust feedback—say, navigation directions, but it could also be the working of a traffic light—based on measurement—in app-to-data cener and car-to-traffic light communication. It is also to say that we can only understand what is going on in that setting, if we take into account its infrastructural extension, spanning field and lab, street and data center.[19] The test situation (and correspondingly the test setting) does not just arise in the street, but across the street and the data centers it connects to.

But such computationally enabled test environments also differ from the classical, controlled experiment. In a real-world test environment, the propensity of the setting to inform, inflect, or influence the social phenomena that unfold within them is considered—by the engineers of the social—a *positive* feature. By contrast, in the philosophy of science pertaining to the classical experimental model, it was considered *a sign of weakness* if you needed to modify environments in order to induce the phenomenon under scrutiny. Such a brittle, poor, inauthentic test was assumed to lack robustness. By contrast, in the setting of the real-world experiment, the more the environment and the entities that constitute it have the proven capacity to influence and modify the behavior of the entities inside them, the more productive, the more successful, from a scientific and engineering point of view, it

is considered to be. As Jonathan Bach (this volume) writes in his essay on the Chinese social credit system, it is a "feature not a failure."

For these reasons we argue that new forms of testing are not occurring within settings so much as they are testing the settings. To figure out what modifications of the smart environment (targeted route redirections, changing settings of traffic lights) enable experimental operations upon societal phenomena (traffic flow, mobility, but also more ephemeral phenomena like trust), of producing the desired modification in the phenomenon under scrutiny (less congestion, increased use of alternative means of transport), settings are continuously modified and continously test users or passers-by. The test-ing settings of experimentation beyond the laboratory, then, are emphatically not the settings that you, the user, can adjust and tweak on your user interface, but the macro settings that are specific to the testing environment itself. Which is also to say, the curation of such test environments in society is intimately connected to the development that we captured hereabove in the slogan: "engineering has moved into the social." As tech industries have overseen an unprecedented extension of infrastructures for data capture, analysis, and feedback across environments in society, in the form of social media, smart transport systems, digital payment, and so on, an extensive environment has become available to scientists and engineers who work with the industries not only to research social life—what is trust?—but also for intervening in it—how to enhance trust?

6 | A CRITICAL MOMENT FOR SOCIOLOGY AND THE SOCIOLOGY OF TESTING

These new developments concerning test settings raise fundamental questions that will need to be addressed by the new sociology of testing. Most crucially—but also most challengingly—we must address the implications for our understanding of "the social": as we detect a shift from tests in settings to testing settings, should we move from studying testing in its social contexts to analyzing the ways in which testing puts social life to the test? We believe the contributions to the Special Issue convincingly make clear that a good place to start is to investigate how testing in society gives rise to test-ing situations: tests do not just render phenomena know-able and action-able, they put social relations at stake (Robinson, this volume; Tironi, this volume). In undertaking such investigation, a new sociology of testing can build on work in pragmatist sociology which focuses on the analysis of *critical moment*s (Boltanksi & Thévenot, 1999; Hutter & Stark, 2015) and *trials of explicitness* (Muniesa & Linhardt, 2011; see also Guggenheim, 2014; Marres, 2012), the idea that when habitual ways of doing get interrupted in social life, whether by accident or as a consequence of deliberate disruption, social actors are prompted to articulate their attachments and relations. These moments in which social actors are put to the test are distinctive as "problematic situations."[20] However, in today's total test environments, because test settings are infrastructurally configured, and because testing can be taking place without one's knowledge at any time, critical moments come about in a different way: they arise from practices of testing and counter-testing in diverse modalities—research, art, journalism—as part of an ecology of testing. Furthermore, as the creation of test environments in society is often done with the explicit purpose of governing and influencing social life, a sociology of testing must come to terms with the possibility that the fundamental "logics of testing" that pragmatist sociologists are trained to detect in situations do not necessarily present a constitutive property of social life, but are materially, technically, and politically inflected by exogenous, interested agencies, like engineering, or "strategic niche management," or experimental governance.

The second question the new sociology of testing will need to address concerns the concept of *proxy* that has been so important in STS studies of testing. In Trevor Pinch's canonical paper (Pinch, 1993; see also Downer, 2007), testing is centrally concerned with a problem of representation. In his account, engineering tests are a "proxy" that *stands for* something "out there" in the world. But in real-world testing, the focus is not on the creation of dedicated, controllable test environments away from society, but on the modification of "real-world" settings

in society. If the test setting (as we argued above) is no longer spatially and temporally separate from the environment in society, does this notion of proxy still apply? It seems clear that the problem of the proxy—the problem of reference—is not going to go away, even as engineering approaches to knowing society are increasingly prominent today. But it might mean that the new forms of testing are neither proxies that are from the outset designed to *stand for* something "in society" nor critical moments that already *stand out* as specific situations as part of these situations. Testing situations will have to be analyzed as unfolding across settings.

Thirdly and finally, it seems likely to us that the study of testing in society will confront sociologists with a choice: what position should we as sociologists take on the new forms of testing and the engineering of the social? We might be tempted to conclude that the very features that make real-world test environments attractive from the standpoint of social engineering—the possibility to impact society through the creation of testing environments—is what renders these environments utterly unsuitable as a site of enquiry for sociology. It signals that insofar as "social life" can be said to occur in test environments, it is inauthentic, an artefact of engineering. It is true that what goes on in test environments is not representative of all of society. But we misunderstand test environments in society if we approach them in "proxy" mode (if we consider them from the standpoint of the field test, and view it as a bastard version of the controlled experiment, and end up saying: "this is not society, society is over there"). Today's test environments are key sites where forms of life, forms of experience are defined, contested, and indeed, tested today.

What, then, is the other option? A sociology of testing could make it its purpose to specify the new methodologies of social engineering from a critical, reconstructive perspective (Entwistle & Slater, 2019; Marres et al., 2018). Such an approach would not go along with narrow framings of the object and objectives written into testing settings: social life does not just unfold inside the test environment (what is trust on Facebook?), the test environment is very much part of it (we should ask: what does trust *become* on Facebook?). Tests do much more than reveal the properties of environments and phenomena unfolding inside them, *tests* operate on social relations, they may reveal capacities and be deployed to hide them. "Testing" is then the name of a fundamental social effect, to be specified by the new sociology of testing. This is what defines situations sociologically speaking—that *they are testing*. But, whatever the answers to our three questions above, one thing seems clear to us: It is in addressing these and similar theoretical and empirical challenges that sociology faces a critical moment in which it is put to the test.

ACKNOWLEDGMENTS

Work for this paper was supported by an Advanced Research Grant from the European Research Council, grant agreement no. 695256. We are grateful to Pablo Boczkowski, Hjalmar Bang Carlsen, James McNally, Elena Esposito, Jonathan Bach, Fabian Muniesa, Trevor Pinch, Joris van Hoboken, and Michael Guggenheim, for comments, criticisms, and suggestions.

DATA AVAILABILITY STATEMENT

This article is based on a dicussion of academic literature and reports available in the public domain.

ORCID

Noortje Marres https://orcid.org/0000-0002-8237-6946
David Stark https://orcid.org/0000-0003-2435-9619

NOTES

[1] With thanks to Willem Schinkel who brought this quote to our attention in his closing remarks at the workshop "Put to the Test: Critical Evaluations of Testing," Warwick in London, December 10 and 11, 2018.

[2] We use the term "laboratory" loosely in this essay, as referring to dedicated, contained sites where controlled conditions have been created for the detection, monitoring and analysis of defined phenomena.

[3] Ronnell (2007), for example, points to tests as the quintessential element of modernity, signaling the drive to move beyond known limitations.

[4] Whether this development is *really new*, or a continuation or re-activation of previous developments, is a question for the new sociology of testing to answer. In this essay we examine the intuition that existing approaches in the sociology of testing leave us under-equipped to analyze testing in society, and formulate questions that a new sociology of testing will need to address.

[5] MacKenzie (1989) draws on the holistic philosophy of science of Pierre Duhem to make the point that tests are "always open to challenge" (p. 430). In his account, it is the job of the sociology of testing to explain why they are not always challenged in practice. "Beliefs about what constitutes adequate testing are conventional, but that does not make them any less significant" (p. 430). Here, as in the French convention school, the preoccupation is with how tests produce legitimate results.

[6] His article begins with an evocative account of the consequences of ballistic missile testing in situ, citing at length a journalistic description of Kwajalein, one of the Marshall Islands that "[the Pentagon] cleared the inhabitants off … and they are now crowded on a tiny speck about two miles north of Kwajalein Island. … without a reliable water supply, without proper medical care, and even without sufficient room to bury their dead…" (p. 409).

[7] ICBM testing—the production of accuracy estimates for intercontinental ballistic missile trajectory predictions at the Vandenberg Air Force base in California—fit both meanings of "hard." Estimates and equations were the hard stuff of engineering, and their isolation from "society" made them a hard (i.e., difficult) case in which to discover the social.

[8] Although a different terminology from that of SSK, this ANT vocabulary nonetheless places limits on our understanding of the role of testing in society today. Concepts like enrollment (Callon, 1984), as well as script (Akrich, 1992), led actor-network theorists and researchers to analyze the role of social actors in technology testing primarily in terms of compliance and resistance. However, as we will discuss below, testing in society today raises different questions: the question is not whether or not social actors are enrolled by means of the test, but which modes of knowing, valuing, and acting get activated as social actors engage with the test: do they, for example, create art works, deploy counter-expertise, or initiate counter-testing, as we will see later on.

[9] Actor-network theory had of course already demonstrated that tests and testing offer powerful instruments for transforming society, by virtue of their displacement—and transportability—from dedicated test settings to environments in society: for example, the displacement of a "smell test" from the perfume laboratory into focus group settings, and/or perfume shops, provides a powerful instrument for creating new products, new experiences, new preferences, and new markets (Latour, 2004; Muniesa, 2014). However, whereas for actor-network theorists, the transformative capacity of test critically depends on the re-production of contained laboratory conditions in social environments (the "laboratization thesis"), we make a different argument: it is through the introduction of tests into distinctively social environments (the street, the city square, social media conversations)—precisely not laboratories, but societal spaces—that science and engineering are today gaining the capacity to extend engineering logics into distinctively social phenomena—trust, interaction in public space, collective behaviour, well-being, and so on.

[10] With thanks to Trevor Pinch for suggesting this formulation.

[11] Michael Guggenheim (2012) argues that the Chicago's School definition of the city as laboratory should be understood metaphorically, as a rhetorical strategy design to make space for social science. We agree here below that the Chicago School approach to the city as laboratory stands out for its lack of interventions in urban environments (they did not intervene to make the city more like a laboratory). However, we go on to develop a different perspective on experiments conducted in urban environments: in our view these experiments precisely do not fit the definition of laboratory that Guggenheim relies on ("a procedure that often results in a space with the properties to separate controlled inside from uncontrolled outside"—p. 101), which defines experiments in terms of the ability to contain phenomena in a controlled environment. This condition of containment, in our view, is lifted, or at least relaxed, in contemporary instances of real-world experimentation.

[12] Sociologist Harry Collins viewed science as an "examplar" and analogy for "all other forms of social and conceptual innovation" (1985, p. 165) where he proposed "science is a representative example of cultural activity …" expressing the hope that "social and political scientists will be able to use his and other modern studies of science to illuminate more general problems" (p. 166).

[13] Similarly, Gross's idea that social experimentation is "different" (Gross & Krohn, 2005) from experimental social science reminds us of Latour's (1988) exclamation in the Pasteurization of society: to transform society with the aid of the vaccine—with the aid of a laboratory—in his account "is completely different" from all other attempts to change society by other means (the law, political mobilization, fashion).

[14] https://jamesbridle.com/works/autonomous-trap-001.

[15] Robinson (2016) discusses a US court ruling that "pregnancy" was not a disease.

[16] Bernhard Rieder, creator of the app for Facebook network analysis and visualization Netvizz, submitted his app for review and was refused access to Facebook data on the grounds that "permissions data must be visibly used within your app" (see Rieder, 2018).

[17] With thanks to Chris Anderson (Leeds University) who suggested this term to us during the December 2018 Put to the Test workshop in London.

[18] They are in some ways similar to what Morgan (2013) calls "society's experiments"—"situations [which] present themselves as if an experimenter has designed a laboratory experiment within the world" (p. 346). Except that in real-world tests, experimenters do in fact design at least some aspects of the test environment in society.

[19] This operation is invoked by Gieryn (2006) in his account of the Chicago School (see the quotation below). However, where continuity between field and lab in Gieryn's account is a methodological accomplishment, in real-world experimentation this operation takes the form of socio-technical and material modifications of connected environments in society so as to create a continuous test environment (for a discussion, also see Marres, 2017, p. 53). "Authors of the Chicago School oscillate between making Chicago (the city) into a laboratory and a field-site. On some occasions, the city assumes the qualities of a lab: a restricting and controlling environment, whose placelessness enables generalizations to 'anywhere,' and which demands from analysts an unfeeling detachment. On other occasions, the same city becomes a field-site, and assumes different qualities: a pre-existing reality discovered by intrepid ethnographers who develop keen personal sensitivities to the uniquely revealing features of this particular place" (Gieryn, 2006).

[20] The notion can be found in John Dewey's emphasis that inquiry—discovery—happens in "troubled, perplexing, trying situations" (Dewey, 1998, p. 140). For Mische and White (1998), a "situation" is a special kind of setting, defined by them as "problematic, … episodes that cast our prescribed roles and trajectories into question" (p. 697). For more on Dewey's notion of inquiry see Stark (2011, pp. 2–9); on methodological situationalism see Stark (2017, 2011, pp. 32, 185–186). Formilan and Stark (this volume) examine critical moments in which electronic music artists probe and test identities.

REFERENCES

Agüera y Arcas, B., Todorov, A., & Mitchell, M. (2018, January 11). Do algorithms reveal sexual orientation or just expose our stereotypes? *medium.com*. Retrieved from https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Akrich, M. (1992). 'The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205–224). Cambridge, MA: MIT Press.

Arms Control Association. (2019, February). The nuclear testing tally. *armscontrol.org*. Retrieved from https://www.armscontrol.org/factsheets/nucleartesttally

Ausloos, J. (2018, April 10). Paul-Olivier Dehaye and the raiders of the lost data. *law.kuleuven.be*. Retrieved from https://www.law.kuleuven.be/citip/blog/paul-olivier-dehaye-and-the-raiders-of-the-lost-data/

Bacharach, Y., Grepel, T., Kohli, P., Kosinski, M., & Stillwell, D. (2014). Your digital image: Factors behind demographic and psychometric predictions fro social network profiles. In *AAAMAS '14: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems.* Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Bellanova, R., & González Fuster, G. (2018). No (Big) Data, no fiction? Thinking surveillance with/against Netflix. *The Politics and Policies of Big Data: Big Data Big Brother.*

Beitelshees, A. L. (2012). Personalised antiplatelet treatment: A RAPIDly moving target. *The Lancet*, *379*(9827), 1680–1682.

Bijker, W. E., & Pinch, T. (2012).Preface to the anniversary edition. In W. E. Bijker, T. Hughes, & T. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. xi–xxxiv). Cambridge, MA: MIT Press.

Boltanski, L., & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory*, *2*(3), 359–377.

Boltanski, L., & Thévenot, L. (2006). *On justification*. Princeton, NJ: Princeton University Press.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295–298.

Boyle, A. (2013, March 12) Gay? Conservative? High IQ? Your Facebook "likes" can reveal traits. *nbcnews.com*. Retrieved from https://www.nbcnews.com/sciencemain/gay-conservative-high-iq-your-facebook-likes-can-reveal-traits-1C8805606

Bruno, L. A. (2003). The bequest of the nuclear battlefield: Science, nature and the atom bomb during the first decade of the Cold War. *Historical Studies in the Physical and Biological Sciences*, *33*(2), 237–260.

Buchmann, S., Lafer, I., & Ruhm, C. (2016). *Putting rehearsals to the test practices of rehearsal in fine arts, film, theater, theory, and politics*. Berlin, Germany: Sternberg Press.

Bulman, M. (2019, April 27). Home Office under investigation after 1,000 suddenly deported over English test cheating claims. *independent.co.uk*. Retrieved from https://www.independent.co.uk/news/uk/home-news/home-office-student-visa-english-language-test-cheat-a8888926.html

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *theguardian.com*. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cage, J. (1973). Experimental music: Doctrine. *Silence*. Middletown, CT: Wesleyan University Press.

Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. InJ. Law (Ed.), *Power, action, and belief: A new sociology of knowledge?* (pp. 196–233). London, UK: Routledge.

Callon, M. (1986).The sociology of an actor-network: The case of the electric vehicle. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 19–34). London, UK: Palgrave Macmillan.

Cambrosio, A., Keating, P., Vignola-Gagné, E., Besle, S., & Bourret, P. (2018). Extending experimentation: Oncology's fading boundary between research and care. *New Genetics and Society*, *37*(3), 207–226.

Cochoy, F. (2004).Is the modern consumer a Buridan's donkey? Product packaging and consumer choice.In K. M. Ekström, & H. Brembeck (Eds.), *Elusive consumption* (pp. 205–228). London, UK: Bloomsbury.

Collins, H. M. (1981). Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science*, *11*(1), 33–62.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London, UK: SAGE.

Consumer Reports. (1959, March). The milk all of us drink and fallout. *Consumer Reports*, *24*.

Davis, T. C. (2007). *Stages of emergency: Cold war nuclear civil defence*. Durham, NC: Duke University Press.

Dehaye, P.-O. (2017, March 3). Cambridge Analytica demonstrably non-compliant with data protection law. *medium.com*. Retrieved from https://medium.com/personaldata-io/cambridge-analytica-demonstrably-non-compliant-with-data-protection-law-95ec5712b61

DeVeaux, S. (1997). *The birth of bebop: A social and musical history*. Berkeley, CA: University of California Press.

Dewey, J. (1954). *The public and its problems*. Chicago, IL: Swallow Press.

Dewey, J. (1998[1933]). Analysis of reflective thinking. *The Essential Dewey*, *2*, 137–144.

Dewey, J. (1998[1925]). *Experience and nature*. New York, NY: Dover Publications.

Downer, J. (2007). When the chick hits the fan: Representativeness and reproducibility in technological tests. *Social Studies of Science*, *37*(1), 7–26.

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441.

Engels, F., Wentland, A., & Pfotenhauer, S. M. (2019). Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy*, *48*(9), 103826.

Entwistle, J., & Slater, D. (2019). Making space for "the social": Connecting sociology and professional practices in urban lighting design. *British Journal of Sociology*, *70*(5), 2020–2041.

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society*, *15*(8), 1248–1365.

Gieryn, T. F. (2006). City as truth-spot: Laboratories and field-sites in urban studies. *Social Studies of Science*, *36*(1), 5–38.

Gieryn, T. F. (2018). *Truth-spots: How places make people believe*. Chicago, IL: University of Chicago Press.

Gross, M. (2009). Collaborative experiments: Jane Addams, Hull House and experimental social work. *Social Science Information*, *48*(1), 81–95.

Gross, M., & Krohn, W. (2005). Society as experiment: Sociological foundations for a self-experimental society. *History of the Human Sciences*, *18*(2), 63–86.

Guggenheim, M. (2012). Laboratizing and de-laboratizing the world: changing sociological concepts for places of knowledge production. *History of the Human Sciences*, *25*(1), 99–118.

Guggenheim, M. (2014). Introduction: Disasters as politics—Politics as disasters. *The Sociological Review*, *62*(Suppl. 1), 1–16.

Hennion, A. (2003). Music and mediation: Towards a new sociology of music. In M. Clayton, T. Herbert, & R. Middleton (Eds.), *The cultural study of music: A critical introduction* (pp. 80–91). London, UK: Routledge.

Hutter, M., & Stark, D. (2015).Pragmatist perspectives on valuation: An introduction. In A. B. Antal, M. Hutter, & D. Stark (Eds.), *Moments of valuation: Exploring sites of dissonance* (pp. 4–16). Oxford, UK: Oxford University Press.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Knorr-Cetina, K. D. (1981).The micro-sociological challenge of macro-sociology: Towards a reconstruction of social theory and methodology. In A. V. Cicourel & K. D. Knorr-Cetina (Eds.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies* (pp. 1–47). Boston, MA: Routledge & Kegan Paul.

Kosinski, M., Stillwell, D. J., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of USA: US National Academy of Science*, *110*(15), 5802–5805.

Lambiotte, R., & Kosinski, M. (2012). Tracking the digital footprints of personality. *Proceedings of the IEEE*, *102*(12), 1934–1939.

Latour, B. (1993). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1996). *Aramis, or the love of technology*. Cambridge, MA: Harvard University Press.

Latour, B. (2004). How to talk about the body? The normative dimension of science studies. *Body & Society*, *10*(2–3), 205–229.

Laurent, B., & Tironi, M. (2015). A field test and its displacements. Accounting for an experimental mode of industrial innovation. *CoDesign*, *11*(3/4), 208–221.

Lezaun, J., Muniesa, F., & Vikkelsø, S. (2013). Provocative containment and the drift of social-scientific realism. *Journal of Cultural Economy*, *6*, 278–293.

MacKenzie, D. (1989). From Kwajalein to Armageddon? Testing and the social construction of missile accuracy. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. 409–436). Cambridge, UK: Cambridge University Press.

Marres, N. (2012). *Material participation: technology, the environment and everyday publics*. Basingstoke, UK: Palgrave.

Marres, N. (2017). *Digital sociology*. Cambridge, UK: Polity Press.

Marres, N., Guggenheim, M., & Wilkie, A. (2018).Introduction: From performance to inventing the social. In N. Marres, M. Guggenheim, & A. Wilkie (Eds.), *Inventing the social* (pp. 17–37). Manchester, UK: Mattering Press.

Mills, M. A. H., & Tkaczyk, V. (in press). *Testing hearing: The making of modern aurality*. Oxford, UK: Oxford University Press.

Mische, A., & White, H. (1998). Between conversation and situation: Public switching dynamics across network domains. *Social Research*, *65*(3), 695–724.

Mody, C. C., & Lynch, M. (2010). Test objects and other epistemic things: A history of a nanoscale object. *The British Journal for the History of Science*, *43*(3), 423–458.

Morgan, M. S. (2013). Nature's experiments and natural experiments in the social sciences. *Philosophy of the Social Sciences*, *43*(3), 341–357.

Muniesa, F. (2014). *The provoked economy: Economic reality and the performative turn*. Routledge.

Muniesa, F., & Linhardt, D. (2011). Trials of explicitness in the implementation of public management reform. *Critical Perspectives on Accounting*, *22*(6), 550–566.

Murphy, M. (2006). *Sick building syndrome and the problem of uncertainty: Environmental politics, technoscience, and women workers*. Durham, NC: Duke University Press.

Neff, G., & Stark, D. (2004). Permanently beta: Responsive organization in the internet era. In P. N. Howard & S. Jones (Eds.), *Society online: The internet in context* (pp. 173–188). Thousand Oaks, CA: SAGE.

Palmer, V. M. (1928). *Field studies in sociology: A student's manual*. Chicago, IL: University of Chicago Press.

Pinch, T. (1993). "Testing—One, Two, Three… Testing!": Toward a sociology of testing. *Science, Technology, & Human Values*, *18*(1), 25–41.

Popper, K. (1961). *The poverty of historicism* (3rd ed.). New York, NY: Harper & Row.

Popper, K. (2002). *The logic of scientific discovery*. London, UK: Routledge.

Rabehariosa, V., & Bourret, P. (2009). Staging and weighting evidence in biomedicine: Comparing clinical practices in cancer genetics and psychiatric genetics. *Social Studies of Science*, *39*(5), 691–715.

Reiss, L. Z. (1961). Strontium-90 absorption by deciduous teeth. *Science*, *134*(3491), 1669–1673.

Rieder, B. (2018, August 11). Facebook's app review and how independent research just got a lot harder. *thepoliticsofsystems.net*. Retrieved from http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/

Robinson, J. H. (2016). Bringing the pregnancy test home from the hospital. *Social Studies of Science*, *46*(5), 649–674.

Ronnell, A. (2007). *The test drive*. Champaign, IL: University of Illinois Press.

Seaver, N. (2019). Knowing algorithms. In J. Vertesi & D. Ribes (Eds.), *Digital STS* (pp. 412–422). Princeton, NJ: Princeton University Press.

Stannard, J. N. (1988). *Radioactivity and public health: A history*. Washington, DC: Office of Scientific and Technical Information.

Stark, D. (2011). *The sense of dissonance: Accounts of worth in economic life*. Princeton, NJ: Princeton University Press.

Stark, D. (2017). For what it's worth. *Research in the Sociology of Organizations*, *52*, 383–397.

Thrasher, F. (1927). *The gang: A study of 1313 gangs in Chicago*. Chicago, IL: University of Chicago Press.

Tolman, F. (1902). The study of sociology in institutions of learning in the United States. II. *American Journal of Sociology*, *8*(1), 85–121.

UNSCEAR. (2000). *Report to the general assembly (Annex C: Exposures from man-made sources of radiation*. Retrieved from http://www.unscear.org/docs/reports/annexc.pdf

van de Poel, I., L. Asveld, & D. C. Mehos (Eds.). (2017). *New perspectives on technology in society: Experimentation beyond the laboratory*. London, UK: Routledge.

Vertesi, J. (2015). *Seeing like a rover: How robots, teams, and images craft knowledge of mars*. Chicago, IL: University of Chicago Press.

Wan, Y., & Kosinski, M. (2017). Deep neutal networks are more accurate than humans at detecting sexual orientation from facial images. *PsyArXiv* preprint. Retrieved from https://osf.io/zn79k/

Watkins, E. (2001). Radioactive fallout and emerging environmentalism: Cold war fears and public health concerns, 1954–1963. In G. E. Allen & R. M. McLeod (Eds.), *Science, history and social activism: A tribute to everett mendelsohn* (pp. 291–306). Alphen aan den Rijn: Kluwer.

Woolgar, S. (1990). Configuring the user: The case of usability trials. *The Sociological Review*, *38*(Suppl. 1), 58–99.

Wu, T. (2017). *The attention merchants: The epic scramble to get inside our heads*. New York, NY: Penguin Random House.

Yaneva, A. (2005). Scaling up and down: Extraction trials in architectural design. *Social Studies of Science*, *35*(6), 867–894.

How to cite this article: Marres N, Stark D. Put to the test: For a new sociology of testing. *Br J Sociol*. 2020;71:423–443. https://doi.org/10.1111/1468-4446.12746